AN INVESTIGATION OF THE POTENTIAL OF COMPONENT ANALYSIS FOR WEATHER CLASSIFICATION*

WALTER I. CHRISTENSEN, JR.** and REID A. BRYSON
Department of Meteorology, University of Wisconsin, Madison, Wis.

ABSTRACT

Selected hourly surface observations from Madison, Wis. and Minneapolis-St. Paul, Minn. are used as basic data for a series of analyses to determine the feasibility of establishing weather classifications. Component analysis (factor analysis) is applied to a sample of January data for Madison to reduce the number of variables needed to suitably describe each day meteorologically and to create orthogonality among these new variables. With these results as the design matrix in regression analysis, a mathematical model for each day is constructed and each day is compared to all other days in order to classify similar days into distinctive weather types. Every day within each class is compared with the synoptic situation for that day to establish whether these types form a reasonable synoptic pattern. The temporal and spatial validity of these newly found weather types is tested by applying the foregoing results to an independent January sample for Madison and an independent January sample for Minneapolis-St. Paul. The basic analytic techniques are then applied to a Madison July sample. Specifically, the results indicate that the elements of a meteorological observation may be expressed by a smaller number of independent components that agree with our knowledge of dynamics; and these newly created components may be applied in a multivariate analysis to establish distinctive weather types. These weather types are synoptically reasonable and their distribution about the usual pattern of Highs and Lows strongly resembles cloud models and photographs from satellites.

1. INTRODUCTION

Many standard meteorological variables are interdependent, either physically or statistically. This study is an investigation of some techniques for reducing the statistical dependence of these variables, and an inquiry into the question of whether the number of variables required to characterize the weather of a day might be reduced, while still accounting for an acceptable fraction of total variance. Clearly, the ultimate reduction of the number of variables would be to a single weather type. The method to be employed for the removal of statistical interdependence involves the transformation of the matrix of correlation coefficients of the original variables into a diagonal matrix. The reduction of the number of variables, essentially by elimination of redundancy, will be accomplished by component analysis. Finally, by expressing the weather of each day in terms of a linear model using the new independent variables, the days will be grouped into types by the Lund [6] method, and the synoptic logic of these types will be examined.

Total variance may be expressed as the sum of common variance, unique variance, and error variance (Harman [5], p. 15). If we assume that error variance is "pure" error, with an expected value of zero, the remaining variance components are common and unique. Common variance is ascribed to common factors (factors involved in more than one variable in a group and thus redundant), and unique variance is ascribed to the unique factor (a factor involved in a single variable of a set). "Common factors are necessary to account for the intercorrelations among the variables while each unique factor represents that portion of a variable not ascribable to its correlations with the other variables in the set." (Harman [5]). As a simple illustration, one may take the case of temperature and dew-point depression in which the minimum correlation coefficient between them is $r=0.707$ if one assumes equal variances for temperature and dew point (Brooks and Carruthers [2], p. 227). Then, if one could forecast the temperature precisely, he would also account for a minimum of 50 percent of the variance of dew-point depression. On the other hand, an error in temperature forecast would also be propagated into the dew-point depression. In more general terms, any error in common variance propagates into all elements of the group.

In the temperature-dew-point-depression example given above, the dependence of the two variables is physical in nature and may be assessed a priori. However, many meteorological parameters are related in ways so complex that the physical cause and effect linkages are not yet clear. For these cases, an a posteriori statistical or climatological relationship may be used pending the complete elucidation of the cause and effect chain. It is with this climatological association of synoptic situation and meteorological parameter that this study is concerned.

It is well known that neither pressure, temperature, nor any other single element of a meteorological observation

*The research for this study was partially supported by National Science Foundation Grant GP-444.

**Current address: Joint Meteorological Satellite Programs Office, Headquarters, U.S. Air Force, Washington, D.C.

will completely describe the "weather" or characterize synoptic situations or "air masses" [1] uniquely. Thus, it would appear that one should explore the possibility of using weather element complexes [2] to establish whether independent distinctive combinations exist which can be associated either with synoptic situations, or "air masses," or both.

One can imagine numerous combinations of weather elements but it is desirable to select only meaningful combinations that will reduce the data package to a manageable size. Additionally, one may further impose the restriction that these complexes must be statistically independent. One method for accomplishing this is component analysis. With this type of analysis, as will be demonstrated, it is possible to get a small number of independent variables which are in unique combinations that are physically and synoptically reasonable. Some of these may be divided into basically either "air mass" properties or synoptic situation properties. After these weather element complexes have been established, they may be combined into a linear mathematical model for each day of the sample to describe the "weather." By comparing each day with each other day in the sample by a technique similar to the method used by Lund [6], similar days may be grouped into distinctive weather types.[3] Each day within each type may be compared to the synoptic situation for that day to test whether these types are related to distinct synoptic patterns.

2. DATA

The basic data used in this study are official U.S. Weather Bureau hourly observations for Madison, Wis. (MSN) and Minneapolis-St. Paul (MSP) as recorded on USWB Data Card No. 684207. These original variables are assumed to constitute a description of the "weather", but many are redundant and interrelated. Because of this interrelationship, each day was described meteorologically as shown in table 1. This representation is not the only option available and it might not be the optimum choice. If one had elected to use all of the available data, the data-handling problem would have been prohibitive and the persistence associated with many elements of hourly observations would have created excessive repetition. On the other hand, any lower frequency than two observations per day could not represent any diurnal variations. Hence, for the purposes of this study the frequency of two per day (0000 and 1200 LST) was considered acceptable.

Variables 1b/d, 2b/d, 3b/d, and 4b/d are dichotomous variables and were treated as occurrence or non-occurrence by entering a "1" or "0" respectively. Variables 6b/d, 7b/d, and 13b/d were modified from the original form. The wind components were computed from the "wind direction" and "wind speed". Variable 13b/d resulted from testing "cloud type" against commonly defined low clouds. That is, in column 58, entries 1–5 are F, ST, SC, CU, and CB, while entries 6–9 are AS, AC, CI, and CS. If 1–5 appeared, the value in column 57 (amount of low clouds) was used; if 6–9 appeared, zero was entered in column 57. The values for 14b/d were computed from

$$\mathrm{MIHI}=\frac{\mathrm{T}-\mathrm{L}}{10-\mathrm{L}} \tag{1}$$

where

MIHI = amount of middle and/or high clouds in tenths
T = total sky cover (TOTSKY) in tenths
L = amount of low clouds (LOW) in tenths.

The values for 9b/d are the 3-hr. pressure change preceding midnight and noon respectively.

TABLE 1.—*Abbreviated form of the original meteorological variables used as the input data for reduction. The "b" and "d" suffixes represent observations in the same variable at midnight and noon respectively.*

Variable		Description
0000 LST	1200 LST	
1b	1d	Thunderstorm or tornado (TSTM).
2b	2d	Liquid precipitation (LPRECIP).
3b	3d	Frozen precipitation (FPRECIP).
4b	4d	Obstructions (OBSTNS).
5b	5d	Visibility (VSBY).
6b	6d	u-component (UCOM).
7b	7d	v-component (VCOM).
8b	8d	Station pressure (PPP).
9b	9d	3-hr. pressure change (APP).
10b	10d	Dry bulb temperature (TDRY).
11b	11d	Wet bulb temperature (TWET).
12b	12d	Total sky cover (TOTSKY).
13b	13d	Amount of low clouds (LOW).
14b	14d	Estimated amount middle and/or high clouds (MIHI).
15b	15d	Amount opaque (OPAK).

3. METHODS

There were four basic data groups analyzed in this study. Four years of January (1955–58) data for MSN and five years of July (1954–58) data for MSN were subjected to three analytical techniques which are designated as component analysis, regression analysis, and objective grouping. The MSP January (1955–58) data and the MSN January (1959) data were used as independent data samples to test the spatial and temporal applicability of the results of the analyses performed on the MSN January (1955–58) data. The following description of the methodology is presented in terms of the analyses of the MSN January (1955–58) data; but since the MSN July data were treated analogously, this general methodology can be applied to the July data by merely changing the appropriate numerical values.

A. COMPONENT ANALYSIS

This technique is similar to factor analysis; but when unities are found in the principal diagonal of the matrix,

[1] "Air masses" as used in this study are not limited to the classical number, in order to leave open the possibility that within each classical division there may be distinguishable subtypes.

[2] Weather element complexes may be likened to the following definition from Shulman and Bryson [7] as they defined a complex climatic variable: "Evaporative stress is a complex variable which considers the combined effects of temperature, humidity, and wind speed."

[3] A weather type is defined as a situation in which the ensemble of meteorological observations exhibits a pre-determined degree of internal consistency, such that these observations can be combined into a distinctive class, which differs significantly from other such classes.

the method should be called component analysis. In keeping with this definition, the analysis performed herein is component analysis, and it was used to provide a smaller number of hypothetical variables from the original data.

Component analysis has not been used extensively in meteorological research, and then mostly in the last decade. White et al. [10] used component analysis for reducing and selecting independent variables in regression analysis as a means of obtaining efficient statistical forecast equations. Aubert et al. [1] employed component analysis to isolate a small number of factors from a complex of variables which would account for most of the variance of a predictand. Grimmer [4] used component analysis to filter fields of 30-day surface temperature anomaly in terms of sets of patterns specific to each month. Steiner [8] also applied component analysis to complexes of climatological elements for selected stations in the United States, and he applied his new components to a multivariate statistical approach for establishing a regionalization of the climate of the United States.

For January, 124 days with observations of 30 variables were available. From these data, in standardized form, the (30×30) symmetric matrix of correlation coefficients of the 30 variables was formed and component analysis was performed on this (30×30) matrix. With a cut-off value of "1" for the eigenvalue [4], component analysis reduced the (30×30) matrix to a (30×9) matrix which accounted for 80 percent of the total variance. From this matrix of eigenvectors and the diagonal matrix with the eigenvalues in the principal diagonal, the (30×9) factor matrix was formed and it was rotated according to the so-called varimax criterion (Harman [5]). This rotated factor matrix having the 30 original meteorological variables as the rows and the 9 statistically independent components for the columns provided the basis for interpreting the new components as well as forming the design matrix to perform regression analysis on the data.

B. REGRESSION ANALYSIS

When the MSN January data are applied to the following mathematical model (in vector notation), the mathematically predicted values are

$$\hat{\mathbf{y}}=\mathbf{xb} \tag{2}$$

where

$\hat{\mathbf{y}}$=(30×124) matrix of original observations
$\mathbf{x}$=(30×9) design matrix=rotated factor matrix
$\mathbf{b}$=(9×124) matrix of regression coefficients

Or, the model for day i is

$$\hat{\mathbf{y}}_i=b_{i1}\mathbf{x}_1+b_{i2}\mathbf{x}_2+\ \ldots\ +b_{i9}\mathbf{x}_9 \tag{3}$$

where

i=1, 2, 3, . . ., 124
$\hat{\mathbf{y}}_i$=(30×1) column vector of weather for day i
$\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_9$=(30×1) column vectors of the nine components

By applying equation (3), a set of normalized b's were computed for each day.

C. OBJECTIVE GROUPING

From equation (3), one can describe each day separately. The correlation coefficient between day i and day j as described by the nine components would be

$$r_{ij}=\sum_{k=1}^{9} b_{ik}b_{jk} \tag{4}$$

After the matrix of correlation coefficients for all the days in the sample was formed, the method described by Lund [6] was applied and, with $r_0=+0.70$ as the threshold value for selecting types, the days were classified into different weather types. This separation into groups defines the weather types that affected the station and their relative frequencies of occurrence. In order to express the characteristics of each type in terms of the 30 original variables, the mean and interquartile ranges for each meteorological variable for each type were computed. From these data it is possible to describe a weather type in terms of composite values of the 30 variables.

The final step in this analysis was to compare the 1200 GMT synoptic situation with each classified day in order to determine if a reasonable relationship existed between types and synoptic situation.

4. RESULTS

A. MSN JANUARY (1955–58)

1. *Component Analysis.*—When the size of the eigenvalue decreased below "1," further factoring of the matrix of correlation coefficients was terminated. In this manner the dimensions of the original matrix were reduced and it was assumed the remaining variance unaccounted for was unique variance. As a result, 9 independent components were created which accounted for 80 percent of the total variance. With selected numerical values from the rotated factor matrix, table 2 was constructed to summarize the results of component analysis.

The details of the mathematical relationships that exist between the original vairables and the independent components may be found in any standard text on factor analysis (e.g., Harman [5]). In terms of this experiment, the following equations illustrate basic relationships which may be used to convert from one set of variables to the other.

$$\hat{}(10b)=-0.87T_i-0.09u_i+0.11v_i-0.20S_d+0.28K_n +0.08K_h-0.11Z_d-0.11K_d-0.06P_d \tag{5}$$

[4] The eigenvalue of "1" was used in this study since any lower value would represent only about 3 percent of total variance at most.

TABLE 2.—*Summary of the MSN January (1955–58) component analysis results. Numerical values were taken from the rotated factor matrix and represent loadings of the original variables on the components.*

Component number	Symbolic notation	Percent variance accounted for	Meteorological variable number	Description of meteorological variables most heavily loaded on the component	Numerical loadings on the component
1	T_i	16.26	10d	1200 TDRY	−0.91
			11d	1200 TWET	−.91
			10b	0000 TDRY	−.87
			11b	0000 TWET	−.87
			8d	*1200 PPP	.48
2	u_i	8.90	6d	1200 UCOM	−.78
			6b	0000 UCOM	−.75
			8b	0000 PPP	.63
			9d	1200 APP	−.54
			8d	*1200 PPP	.38
3	v_i	9.32	7b	0000 VCOM	−.78
			7d	1200 VCOM	−.76
			9b	0000 APP	.65
			9d	1200 APP	.60
			8d	*1200 PPP	.40
4	S_d	8.92	12d	1200 TOTSKY	−.93
			15d	1200 OPAK	−.88
			14d	1200 MIHI	−.65
			13d	*1200 LOW	−.37
5	K_n	12.09	13b	0000 LOW	.75
			15b	0000 OPAK	.75
			3b	0000 FPRECIP	.75
			12b	0000 TOTSKY	.69
			5b	0000 VSBY	−.62
			4b	*0000 OBSTNS	.48
6	K_h	5.56	14b	0000 MIHI	.90
			12b	0000 TOTSKY	.58
			15b	0000 OPAK	.45
			4b	*0000 OBSTNS	−.24
7	Z_d	9.29	4d	1200 OBSTNS	−.77
			5d	1200 VSBY	.72
			3d	1200 FPRECIP	−.69
			5b	*0000 VSBY	.40
8	K_d	5.35	13d	1200 LOW	−.69
			14d	1200 MIHI	.65
			2b	0000 LPRECIP	−.57
			8d	*1200 PPP	−.18
9	P_d	4.41	2d	1200 LPRECIP	−.93
			6b	*0000 UCOM	.26
Total		80.10			

*Denotes next largest term, *not* used to describe the component but showing the selected cut-off point.

and

$$T_i = 0.00(1b) - 0.23(2b) + 0.02(3b) + \ldots - 0.87(10b) + \ldots - 0.24(15d) \quad (6)$$

or

$$T_i \approx -0.87(10b) - 0.87(11b) - 0.91(10d) - 0.91(11d) \quad (7)$$

It is assumed in equation (5) that each original meteorological variable is completely and linearly determined by the nine components; however, this model does not exclude the possibility that non-linear cause and effect relationships may exist. Furthermore, it should be pointed out that equation (6) is only precise if the original variables (1*b*, 2*b*, etc.) are "ideal" [5] variables. Equation (7) illustrates how one may describe an independent component in terms of those original variables most heavily loaded on the component. Hence, the entries in table 2 indicate those grouped variables which, together, best describe the component in ordinary meteorological terms.

T_i is obviously a temperature-related variable. It represents, for the surface, a point on a Rossby diagram and thus identifies an "air mass". u_i and v_i combine the wind and pressure variables to give a unique form of the wind in which lateral shearing stresses have no importance in describing the "weather." Since many different winds may be observed without necessarily changing "air mass", u_i and v_i are primarily associated with the synoptic situation. S_d is a measure of noon sky conditions and K_n is a measure of midnight sky conditions; both are heavily loaded with middle and/or high clouds. As such, they are mainly related to synoptic situation. K_n, Z_d, and P_d are also dominantly synoptic situation properties, as is K_d. K_d indicates the occurrence of liquid precipitation at midnight along with noon low clouds and an absence of middle and/or high clouds; however, this relationship may not be real. The apparent lack of simultaneous occurrence between LOW and MIHI may be an artifact resulting from the inadequacy of observing clouds from the ground.

In summary, component analysis reduced the number of variables required to explain a significant amount of total variance from 30 interrelated variables to 9 independent components. It isolated those original meteorological variables which, when grouped by the analysis, basically described each component. Even though the separation of the components into "air mass" and synoptic situation is a simplification, some semblance of logical coherence can be noted in each group; however, more stringent division of the components might produce more descriptive separation. For example, u_i and v_i may well be classified as kinematic properties, K_n may be considered as a property that incorporates sky conditions along with the phenomena classed as "present weather" elements in the synoptic code, etc. Of the 80 percent of total variance accounted for by all nine components, about one-fifth was attributed to "air mass" properties and about four-fifths to synoptic situation. The character of a day in MSN in January is primarily related to the synoptic features on that day.

2. *Regression Analysis and Objective Grouping.*—The regression analysis gave one equation for each of the 124 days which provided the *b*'s for the application of equation (4) and the ensuing correlation matrix. An analysis of variance was performed testing the null hypothesis of zero effects due to the nine independent components against the alternative hypothesis that these effects are non-zero. The appropriate test statistic is the *F*-ratio, which is the ratio of the mean square of the explained sum of squares divided by the mean square of the residual sum of squares. This random variable is distributed as a Snedecor *F* distribution (variance ratio distribution) with (9, 19) degrees of freedom. Nineteen degrees of freedom for the residual are available rather than 21 because the two original variables 1*b*/*d* happen to be zero throughout this January sample. The results of testing at several levels of significance showed that all 124 equations exceeded

[5] Harman [5] defines "ideal" variable as "... the variables projected into the commonfactor space."

the 10 percent level of significance; and, in fact, 103 exceeded the 0.1 percent level. Thus, one leans toward the alternative hypothesis which says that the model is reasonable in light of the evidence at least at the 10 percent level of significance.

It may be logically argued that a better way of typing the days would be accomplished by using a Q-mode factor analysis. This possibility was not overlooked but available computer facilities can not invert a matrix as large as (124×124). As a result, the method of typing used herein was selected as the next best way to classify the days.

After applying this method, 107 of the 124 days were classified into 25 groups which were designated as weather types (indicated in following sections by letters of the alphabet). The 17 days that were not typed were examined to determine whether one could call them transitional days among groups. If a logical transitional day is defined as a day preceded by a day of one type and followed by a day of another type in a logical sequence, 11 of the 17 days could be classified as logical transitions Two days were indeterminant because of insufficient continuity (first and last days of the sample) and two sets of two successive days represented reasonable transitions. On the other hand, any of these 17 days may represent rare cases which might have been typed in a larger sample.

Each day classified in a type was compared to the synoptic situation which was observed at 1200 GMT of that day to test whether these types represented true synoptic patterns or whether this method produced types that fell into classes regardless of synoptic situation. A plot of these comparisons is depicted in figure 1 which represents a generalized synoptic chart for January. This chart is schematic but it is patterned after actually observed cases within the sample period. The distances are relative; the shapes and orientations of the pressure systems and the state of maturity of the cyclone are

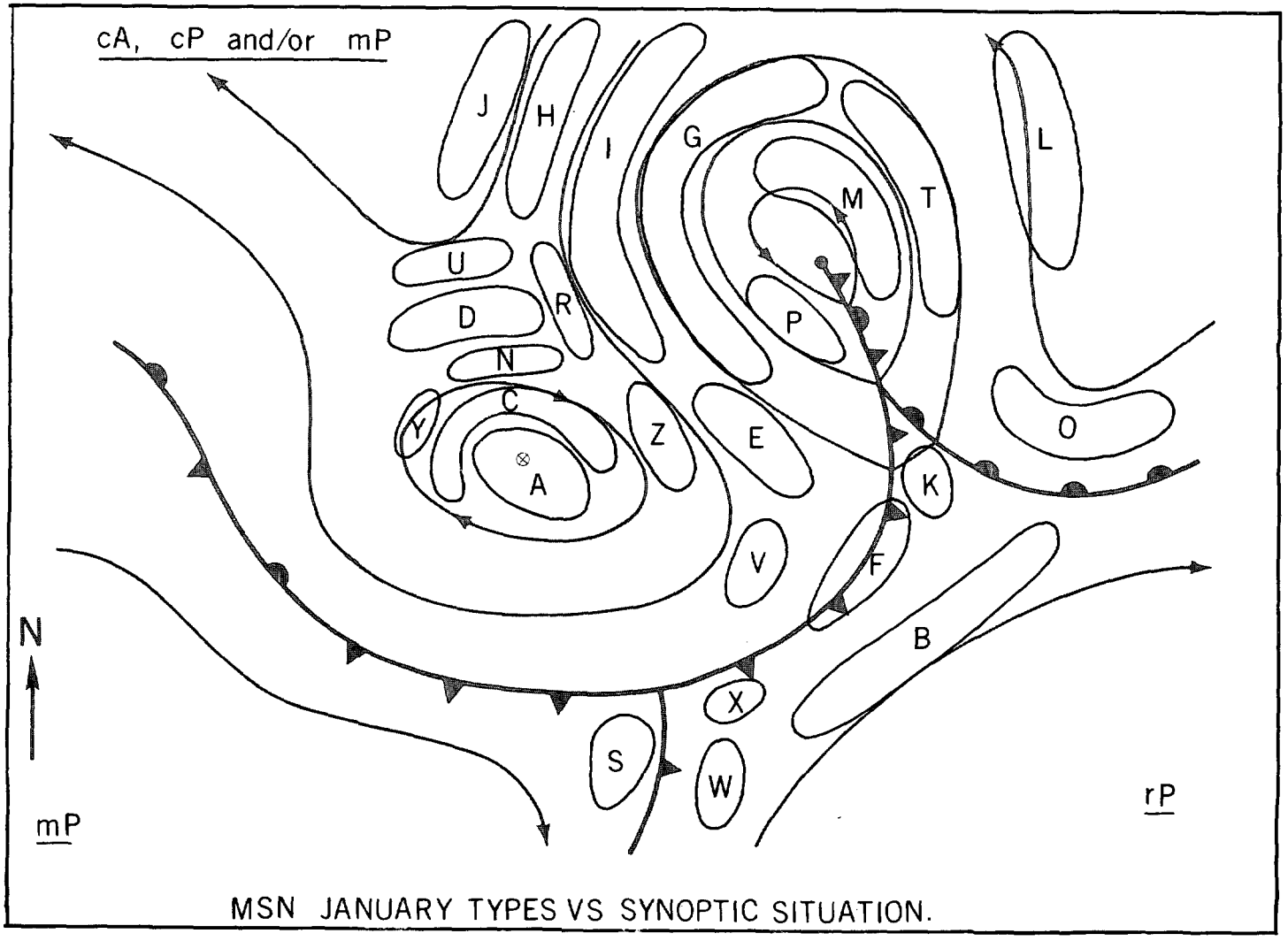

FIGURE 1.—Distribution of Madison (MSN) January types vs. generalized synoptic situation.

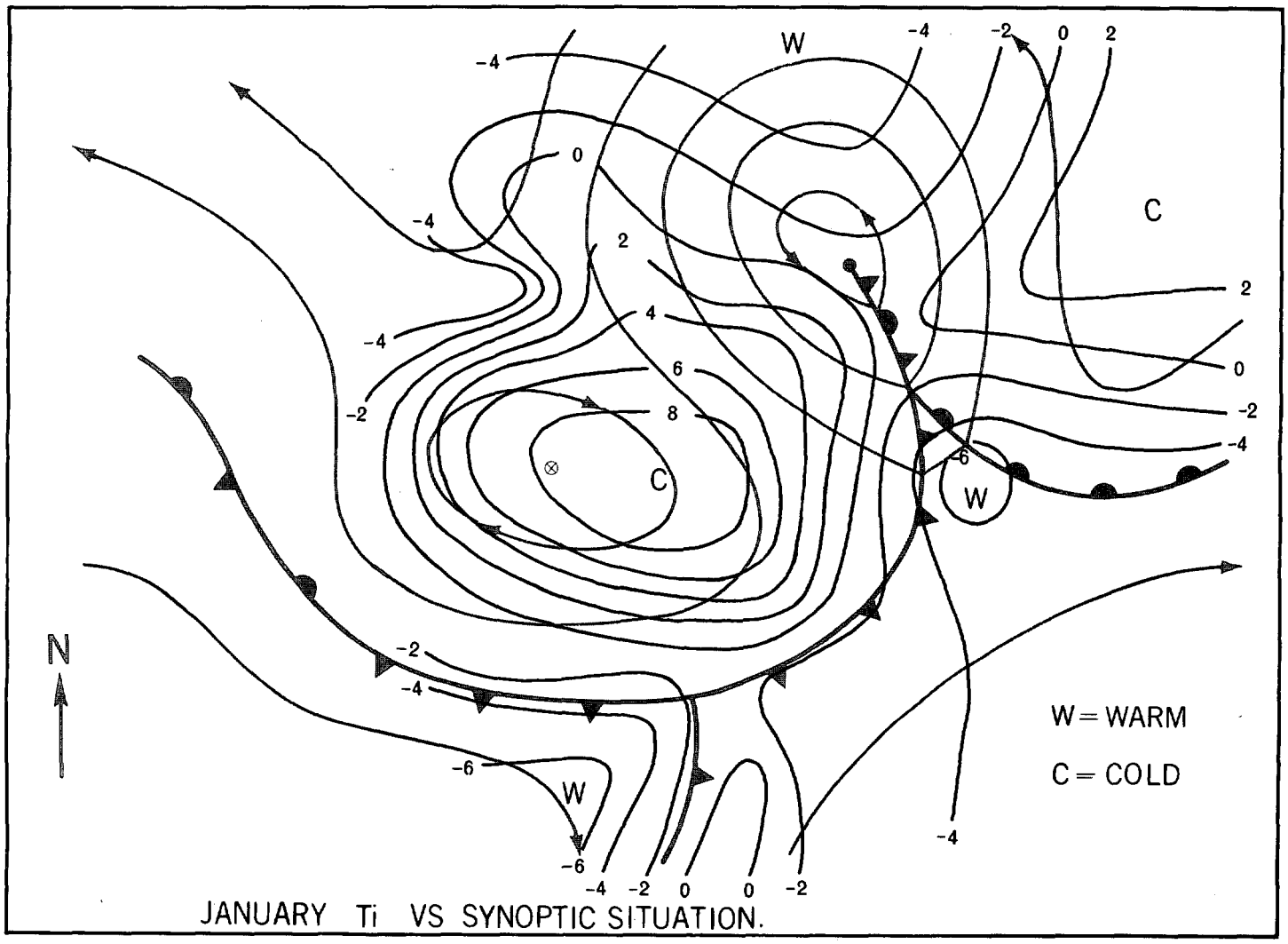

FIGURE 2.—January T_i vs. synoptic situation used in figure 1. Numerical values represent 10 times the value of b_{i1}. Negative values are warm, positive values cold.

generalized in this diagram. The "air masses" are represented as cA (continental Arctic), cP (continental polar), mP (maritime polar), and rP (return polar). The areas to the north of the frontal system as well as the rP air south of the front may be composed of cA, cP, mP, or any combination of these three. The enclosed areas labeled with letters of the alphabet show those areas in which all of the days within that type were observed. The letters represent the associated weather type and for that type MSN would be located within that area. Since, in this study, one cannot deduce the exact transition from one type to another, the boundaries of the types were drawn to indicate the *observed absence of overlap between types*. Allowing for the existence of transition zones, one should view these divisions as the "cores" of the individual types.

Probably the most notable feature of figure 1 is the array of weather types both in the cyclonic and anti-cyclonic areas and the indicated sequence of types as the synoptic features move. The "Norwegian Model", by comparison, emphasizes the precipitation and cloud structure in the vicinity of the Low and its accompanying frontal system. Such a model tells little or nothing about the other elements of the "weather" outside of its area of emphasis. On the other hand, figure 1 indicates weather types in all sectors of the synoptic pattern which also includes more elements than merely clouds and precipitation. In fact, the arrangement of the "cores" shows a striking similarity in shapes and orientations to satellite cloud models and photographs (e.g., Elliott and Thompson [3]).

The importance of the sequence of "weather" as systems move can not be over-emphasized. The knowledge of these sequences adds detail to the forecast. By incorporating the expected movement of the synoptic features as well as the expected changes in intensities of the pres-

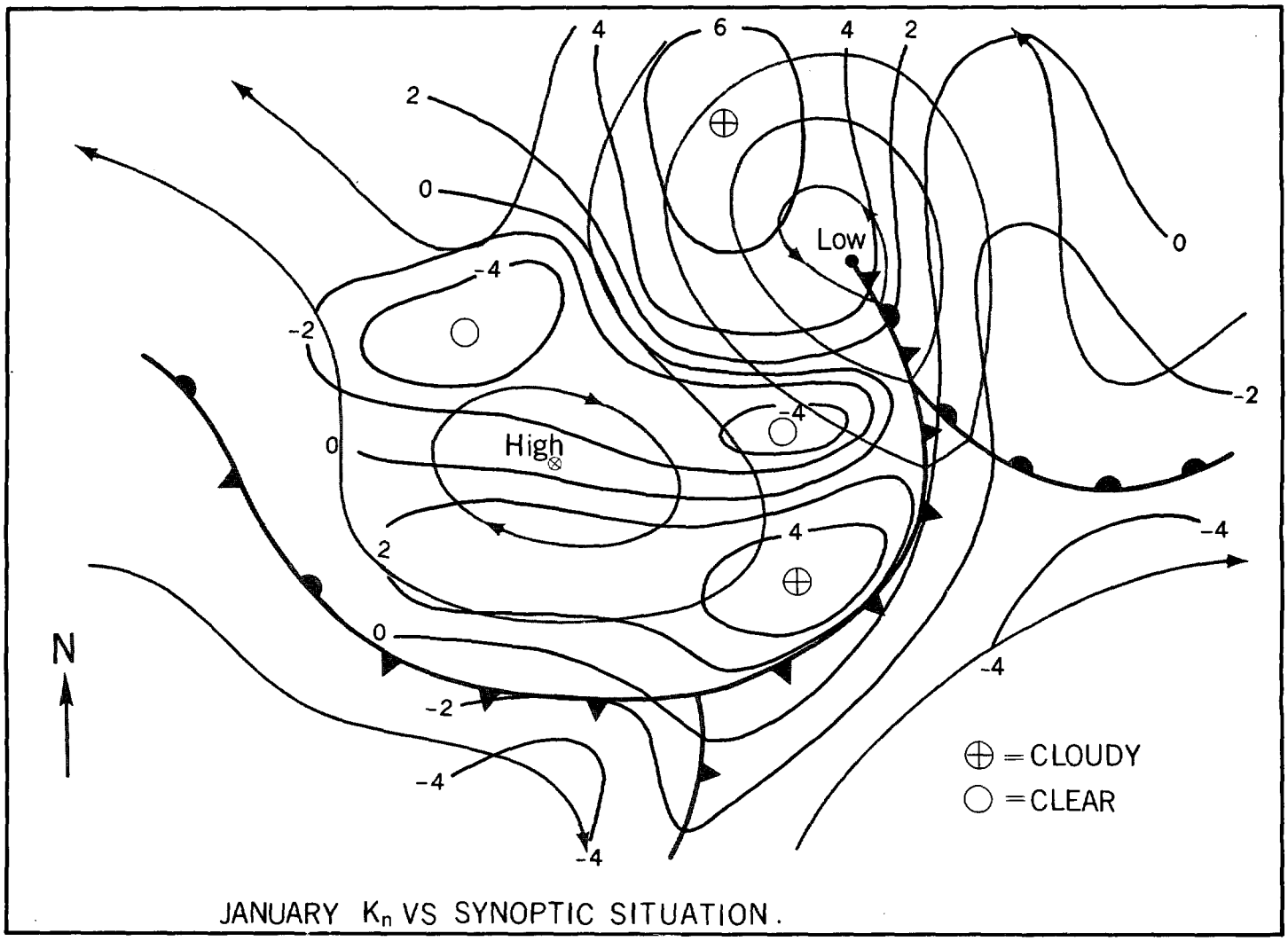

FIGURE 3.—January K_n vs. synoptic situation used in figure 1. Numerical values represent 10 times the value of b_{i5}. Negative values are clear, positive values cloudy.

sure systems, one may be able to establish the details of such a sequence with the aid of figure 1. The following two examples illustrate two different sequences that were observed under two different sets of conditions. The first sequence of T, F, A, and D during the period January 28 through 31, 1957 was associated with a relatively uniform movement of a cold front followed by a cold High and a col area behind the High. By comparison, a sequence of G, H, Z, Z, A, and A was observed during the period January 19 through 24, 1956. This sequence resulted from a Low in the Lake Superior region remaining relatively stationary for the first four days. Cold air was advected into the MSN area and a cold High settled in behind the Low giving rise to very little change in weather type during the latter part of the period. Hence, one should realize that it is not implicit that one type will follow another merely because they are adjacent to each other on the schematic model; rather, the movement, dimensions, and state of maturity of the pressure systems will determine the observed sequence of types. In fact, since these types are defined in terms of a whole day, it is not likely that adjacent types will follow each other because systems normally move at a much higher speed.

In order to test the applicability of using the new independent components for comparison with standard meteorological conditions, the spatial distributions of T_i and K_n versus synoptic situation were analyzed. The results of these analyses are contained in figures 2 and 3 for T_i and K_n, respectively. The synoptic features on each chart are the same as used in figure 1 with the same flexibility in relative conditions.

The spatial distribution of T_i in figure 2 is closely related to what one would expect for a standard temperature measure in January. The colder areas coincide well with the positions of the cold Highs and most of the warm sectors appear where one would expect them. However,

it is noteworthy that this distribution also illustrates very well the warming on the back side of the "Bubble High" and the warming usually associated with the mP fronts in the winter at MSN. Clearly, then, it is reasonable to compare spatial distributions of T_i with other forms of temperature.

The distribution of K_n in figure 3 was selected to illustrate the spatial distribution of a synoptic situation component. K_n is heavily loaded with midnight low cloudiness, frozen precipitation, and visibility, but, for the sake of brevity, maximum and minimum areas on figure 3 are indicated by a cloudy symbol and clear symbol, respectively. The experienced forecaster will easily recognize the distribution of this type of "weather" as being quite realistic and he can easily translate these details into a general forecast. For example, to the northwest of the Low he would expect low ceilings with frozen precipitation and reduced visibility. The minimum just south of the Low lying between two maxima may be recognized by the experienced forecaster. Often the adverse conditions change to fair conditions a short time after cold frontal passage only to be followed by adverse conditions again on the back side of the Low. Obviously, K_n may be easily compared to some of the standard meteorological elements.

B. MADISON (MSN) JULY (1954–58)

The sample for MSN July encompasses one more year of data than MSN January but one day was missing so the total sample contains 154 days.

1. *Component Analysis.*—Again terminating the factoring of the correlation matrix when the eigenvalue decreased to below "1", component analysis reduced the 30 original variables to 10 independent components while accounting for 80 percent of the total variance. Thus, one more component was needed in July to account for the same amount of variance that nine components accounted for in January. The results of the component analysis are summarized in table 3.

T_a is very similar to T_i and, as such, it is basically an "air mass" property. Z_n and B_d are also primarily "air mass" components since they appear to be highly related to low-level stabilities which suggests an association with the "k" or "w" property of an "air mass." The remaining components are quite clearly more related to synoptic situation than to "air mass" except for R_n and R_d. R_n and R_d are measures of midnight thunderstorms with rain and noon thunderstorms with rain, respectively. These components have roots in both "air mass" and synoptic situation. They may be identified with "air mass" in the association of nearly all Wisconsin thunderstorms with maritime tropical air. On the other hand, the "triggering" for nearly all thunderstorms in the Midwest may be related to synoptic situation. Since thunderstorms are not observed each time maritime tropical air influences a station, R_n and R_d will be classified as primarily synoptic situation properties.

Aside from the different number of components for July and January, some other features might be noted. First, the apparent rise in importance of "air mass" properties in July resulted in "air mass" properties accounting for nearly 25 percent of the 80 percent variance accounted for by all 10 components. This indicates that the character of a day in July at MSN may be more related to "air mass" properties than a day in January. Second, the rotation of the factor matrix grouped the original variables similarly for two components in July and January but the other groupings were, in many cases, quite dissimilar.

It is not surprising that the "weather" at MSN in January and July exhibits differences, because it is well known that the synoptic situations are different in January and July. On the other hand, one may note that in both cases the independent temperature components (T_i and T_a) were significantly loaded with the same original variables and, in each case, the temperature component accounted for more variance than any other single component. This similarity is logical if one recalls the relationship of T_i and T_a to the Rossby diagram, the utility of which is well tested. The initial variables of which T_i and T_a are primarily composed are those which, for quasi-constant pressure, define the

TABLE 3.—*Madison (MSN) July component analysis results. Numerical values were taken from the rotated factor matrix and represent loadings of the original variables on the components*

Component number	Symbolic notation	Percent variance accounted for	Meteorological variable number	Description of meteorological variables most heavily loaded on the component	Numerical loadings on the component
1	A_d	9.88	12d	1200 TOTSKY	0.96
			15d	1200 OPAK	.88
			14d	1200 MIHI	.74
			10d	*1200 TDRY	−.40
2	T_a	11.83	10b	0000 TDRY	−.87
			11b	0000 TWET	−.87
			11d	1200 TWET	−.84
			10d	1200 TDRY	−.70
			13d	*1200 LOW	.29
3	Z_n	7.20	4b	0000 OBSTNS	.86
			5b	0000 VSBY	−.80
			13d	*1200 LOW	.47
4	V_d	7.09	9d	1200 APP	.83
			7d	1200 VCOM	−.74
			8b	*0000 PPP	−.47
5	A_n	9.49	12b	0000 TOTSKY	.90
			15b	0000 OPAK	.86
			14b	0000 MIHI	.85
			13b	*0000 LOW	.29
6	R_n	7.54	1b	0000 TSTM	−.87
			2b	0000 LPRECIP	−.82
			13b	0000 LOW	−.60
			15b	*0000 OPAK	−.35
7	R_d	6.89	1d	1200 TSTM	−.84
			2d	1200 LPRECIP	−.82
			5d	*1200 VSBY	.36
8	W_i	7.39	6b	0000 UCOM	.79
			6d	1200 UCOM	.74
			7b	0000 VCOM	.60
			8b	*0000 PPP	−.31
9	B_d	5.61	4d	1200 OBSTNS	−.92
			5d	1200 VSBY	.69
			14b	*0000 MIHI	−.18
10	P_i	7.52	8d	1200 PPP	−.77
			9b	0000 APP	−.74
			8b	0000 PPP	−.62
			7b	*0000 VCOM	.37
Total		80.44			

*Denotes next largest term, *not* used to describe the component but showing the selected cut-off point.

potential temperature and mixing ratio. Since these are physically related to the modification of "air masses", it would be most surprising if they did not appear significant, in combination, at any time of year, at most places on earth. Further similarity between January and July is indicated by S_d and A_d being significantly loaded with the same original variables. Their association with middle and/or high clouds suggests that the occurrence or non-occurrence of this type of cloudiness in both months might be attributed to similar upper-air patterns. In fact, after comparing several days from each sample with the 500-mb. chart for corresponding days, the occurrence of this type of cloudiness was found to be highly related to the proximity of a trough at 500 mb. for both months.

2. *Regression Analysis and Objective Grouping.*—The regression analysis for July was based on one equation (of 10 components) for each of 154 days. The analysis of variance using the previously defined F-ratio as the test statistic revealed that 150 cases exceeded the 5 percent level of significance while the remaining four cases exceeded only the 25 percent level. 134 of the 154 days were classified into 30 weather types. Or, the method typed 87 percent of the days in July as compared to 86 percent in January, indicating that the method is equally applicable in both months for the samples used in this study. With the same definition of a logical transition as used for the January case, an evaluation of the 20 non-typed days revealed 11 logical transitions. Additionally, there were three sets of two consecutive days and one set of three consecutive days which showed reasonable sequences of weather types. Values of the statistical properties of all the variables for all 30 types were computed and they showed narrow limits similar to the January case.

The comparison of the typed days with the synoptic situation is depicted in figure 4. The definition and interpretation of figure 4 are similar to figure 1 with only two exceptions. First, it should be noted that the basic synoptic pattern for July is different from that for January; and second, different "air masses" are listed. Those "air masses" listed for July that appeared in figure 1 have the same definition except for mP. In July, mP air at MSN comes over a variety of trajectories and suffers extensive modification during its course, so the mP for July may be basically defined as air having an initial origin over the Pacific. Three new "air masses" appear in July and they are defined as cT (continental tropical), mT (maritime tropical), and A (air of Arctic, subarctic, or Hudson Bay origin). The absence of cP and cA on the July chart results from there being no real source for these "air masses" in July. The anticyclone to the east of the occlusion may contain any or selected combinations of the indicated "air masses" since the warm front may be displaced markedly northward at times and at other times its surface position is quite vague.

The spatial distribution of the synoptic situation component R_n is depicted in figure 5. This component is most heavily loaded with midnight thunderstorms, liquid precipitation, and low clouds. The areas designated by "L" in figure 5 represent those areas where thunderstorms along with liquid precipitation and low clouds are likely and the areas marked with "U" indicate the unlikely areas. The distribution of R_n allows one to conclude that the occurrence of nighttime thunderstorms at MSN in July is highly dependent upon the proximity of an active frontal system. The further use of this component to relate associated standard meteorological conditions may be beneficial to those working with severe weather forecasting.

5. TESTS OF TYPES ON INDEPENDENT DATA

The results in the previous section seem to indicate that the method used here does, in fact, classify most of the days within the sample into groups which may be associated with synoptic situations. This implication is borne out by figures 1 and 4; but these figures alone may leave some doubt as to the credibility of the types. In order to test whether these types could be applied outside the sample from which they were derived, an independent sample of MSN January (1959) was used. And finally to test the spatial application of the types, the same sample period (1955–58) for MSP January was evaluated.

A. MADISON (MSN) JANUARY (1959)

The data from these 31 days were inserted into equation (2) as the $\hat{\mathbf{y}}$-data, while the data in $\mathbf{x}$ remained as computed from the MSN January (1955–58) data. In other words, the regression analysis was performed using the previous 4 yr. as the causal variables while the 1959 data became the effect variables. The analysis of variance performed on these 31 equations using the previously defined F-statistic showed that all cases exceeded the 2.5 percent level of significance.

By the method described in section 3C, each of the 31 days was compared to each of the 25 predetermined types and again with $r_0=0.70$, 23 of the 31 days were classified into 11 of the 25 types. The 1200 GMT synoptic charts were again consulted to plot the position of MSN with respect to the synoptic features for the classified days on a schematic chart like figure 1. All 23 cases fell into the previously established "core" limits on figure 1, indicating that the components from the previous 4 yr. of data may be used to describe the "weather" adequately in this independent sample; and when comparing these days with the previously derived types, the method does, in fact, classify most of the days in this independent sample into groups which are easily associated with synoptic situations.

B. MINNEAPOLIS-ST. PAUL (MSP) JANUARY (1955–58)

It is well known that not all local effects for surface observations can be completely removed to establish truly standard observations for all stations. After deriving a

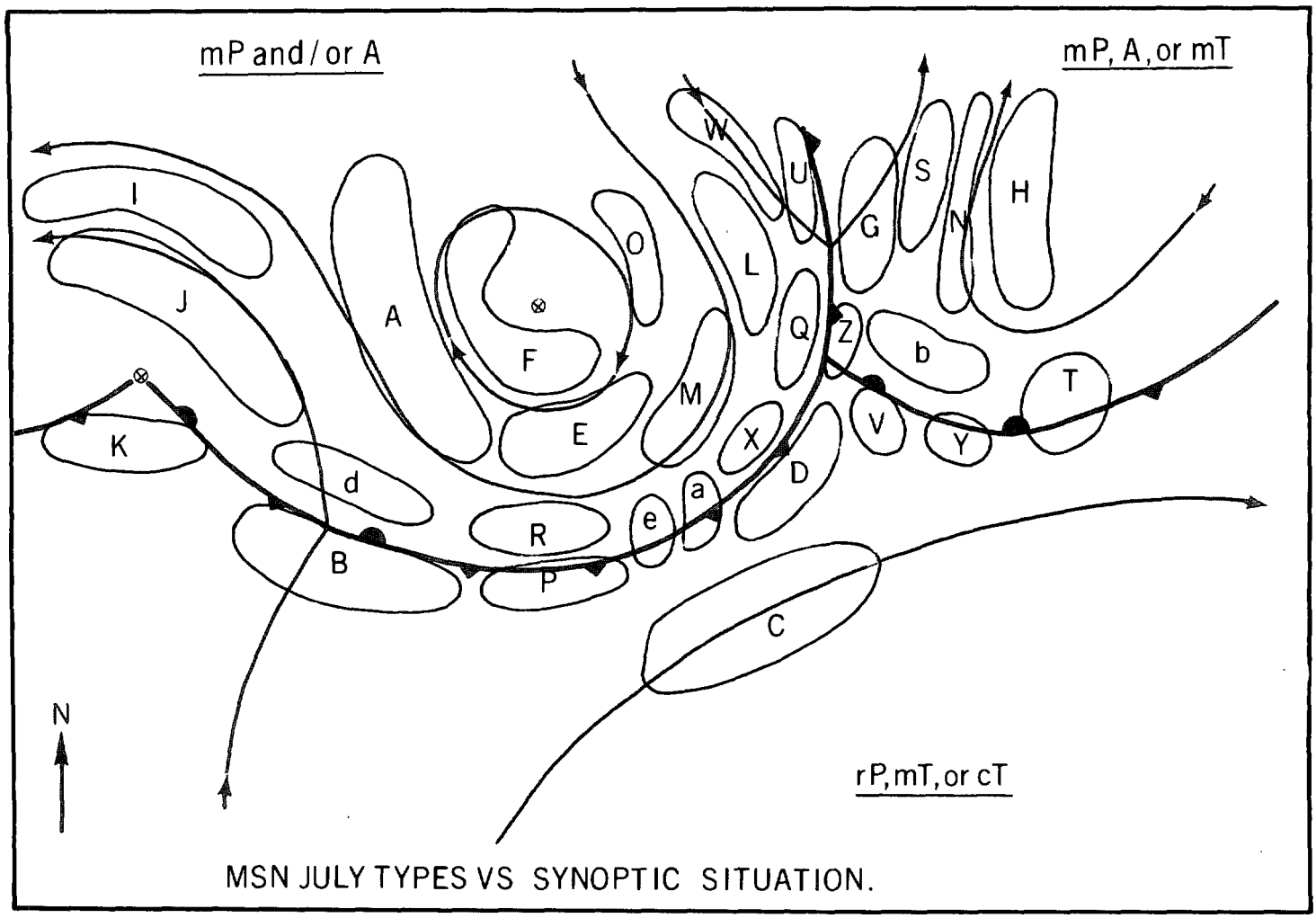

FIGURE 4.—Distribution of Madison (MSN) July types vs. generalized synoptic situation.

set of components from the data for a selected station, one might conclude that these components can be applied only to the station for which they were computed; however, it is possible that a greater generality can be attached to these components. Thus, as a test of the lateral applicability of the nine components derived from the MSN January data (1955–58), the MSP January data (1955–58) were used in equation (2) as effect variables and the nine independent components from the MSN data supplied the causal variables. The analysis of variance using the previously defined F-statistic showed that 110 of the 124 equations exceeded the 5 percent level of significance, 115 exceeded the 10 percent level, and only 2 failed to exceed the 25 percent level. As a matter of comparison, MSN January (1955–58) had 98 percent of the cases exceed the 5 percent level of significance and 100 percent exceed the 10 percent level, while MSP had 89 percent exceed the 5 percent level and 93 percent exceed the 10 percent level. Hence, even though fewer cases exceeded the 10 percent level of significance in the MSP data, it is still quite reasonable to accept the non-null hypothesis for the MSP data.

From the application of the previously defined classification method, 90 of the 124 days were classified according to the 25 types established from the original MSN data. The 34 non-typed days contained two with no continuity (first and last days of a month), three sets of two successive days, two sets of three successive days, one set of four successive days, and twelve single days that showed logical transitions. This greater number of non-typed days at MSP might indicate that the weather systems moved more rapidly in the vicinity of MSP than they did near MSN; and, another possibility is that there may be more "rare" cases or a few "different" types in this MSP sample. The 90 classified days were compared to the 1200 GMT synoptic situation and the results were plotted on a chart like figure 1.

When the MSP results were compared with the MSN

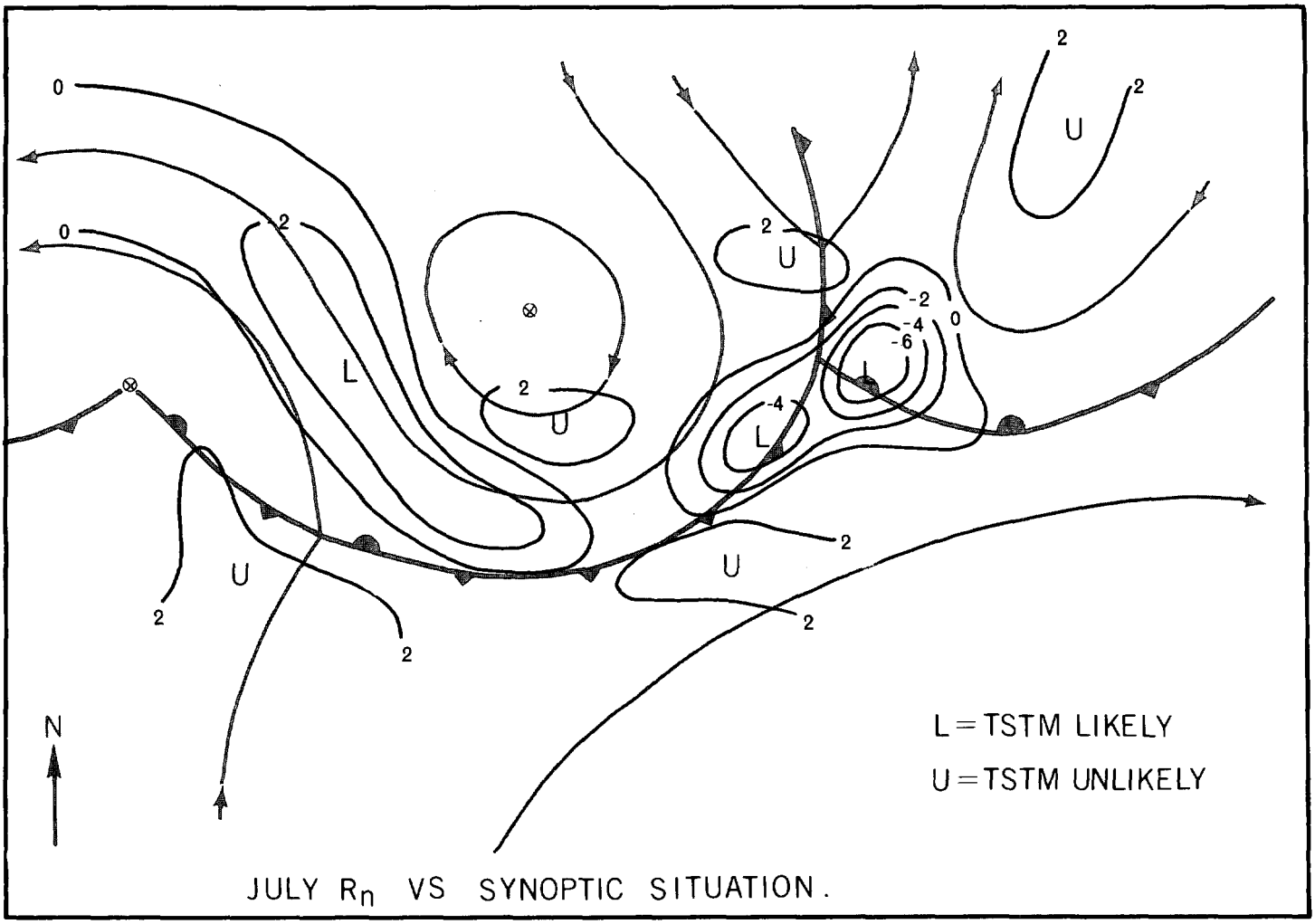

FIGURE 5.—July R_n vs. synoptic situation used in figure 4. Numerical values represent 10 times the value of b_{i6}. Negative values correspond to likely areas and positive values to unlikely areas of thunderstorm occurrence.

results, 90 days (73 percent) were typed for MSP, while 107 days (86 percent) were typed for MSN. Twenty-four of these days were classified as the same for both stations. As one would suspect from figure 1, all of these days that were commonly typed were days belonging to types that have extensive lateral limits. As stated earlier, the lateral limits of the types in figure 1 are neither precise nor is it implicit that they represent the true population of all Januarys. In fact, the incorporation of the MSP data indicates that many of the "cores" shown in figure 1 may occupy more area than illustrated. Comparison of MSP to MSN by type follows.

The MSP data showed 12 A-type days and all of them fell into the same area as the 8 A-type days did for MSN. The greater number of A's for MSP indicates that these cold H gh centers were more frequently near there than near MSN and their movement from MSP either carried them too far from MSN or they moved too fast to allow MSN to be classified as an A-type with the same frequency as MSP. Of the 8 A-days at MSN, 6 were also classified as A-type at MSP indicating that the lateral extent of the A-type is large enough to encompass both stations.

The B's and C's showed good agreement except that the MSP data indicated that the "B-core" may extend farther toward the cold front than previously thought. Also, no significant disagreements were found in types E, J, N, P, T, U, W, and X. Type D showed a marked difference in frequency (3 for MSP and 9 for MSN) and the MSP days appeared to cluster in the western extremity of the "D-core." The MSP data for type F indicated that the lateral extent of the "F-core" from the cold front northwestward is greater than shown for the MSN data.

Four types (G, L, V, and Y) did not appear in the MSP classification. The absence of V and Y is not surprising since each type contained only two cases in the MSN data. The lack of L and G cannot be explained as readily, however. The G-type is characterized by midnight and noon frozen precipitation on the back side of the Low.

The lack of this type at MSP may be due to the trajectory not being over a suitable moisture supply, the lack of a suitable moisture supply close enough to MSP, a weaker circulation about the storm as it passes MSP, or a combination of these. The absence of the L-type at MSP probably results from the lack of storms passing far enough to the west of the station in this particular sample.

Types H, I, O, Z, R, and S for MSP showed that the "cores" occupy larger areas than indicated from the MSN data. The frequencies are in good agreement for all these types except for Z and S. The greater number of Z's for MSP (9) might result from the cold Highs passing close to MSP more frequently than they do at MSN (5). The higher frequency of S's at MSP (6) may result from the mP Highs containing more pure mP air as they influence MSP, while they become more contaminated with other air before they reach MSN (3) and thus gain characteristics that are not representative of the S-type.

Types K and M showed nearly the same frequencies for both stations but the MSP data indicated that these "cores" should be extended through a frontal zone giving evidence that immediately ahead and immediately behind the front, in these relative positions, the "weather" does not change enough to alter the classification.

Essentially, the types established for MSN are applicable to MSP. The foregoing differences are only minor and actually tend to add information about the previously established types rather than detract from their significance. Hence, it appears that one may use a set of components derived from the MSN January data to classify the "weather" into types that have been previously established from the MSN observations. In the upper Midwest where local effects are minimal, the independent components and resulting weather types from one station's data apparently can be applied to other similar stations within a radius of approximately 250 mi.

6. CONCLUSIONS, APPLICATIONS, AND SUGGESTIONS FOR FURTHER RESEARCH

Some of the problems that are basic to weather classification have been attacked in this paper. A solution to the problem of expressing meteorological observations with a minimum number of statistically independent terms was presented using component analysis. This analysis provided a set of components that have the desirable property of orthogonality, a rarity in nearly all meteorological observations. As a result, one can use a mathematical model that is free of interaction terms and possibly apply this to some numerical forecasting technique. The rotation of the factor matrix enables one to interpret the new components in terms of those original variables most heavily loaded on the components.

One may logically think that the results of this classification scheme may be compromised since these samples do not represent sets of independent observations because of the well-known persistence of the meteorological parameters that exists between successive days. In fact, this persistence may not be limited to the individual parameters as some general descriptions of the "weather" may exhibit significant persistence also. For example, a clear, calm, and cold day may very well be followed by another clear, calm, and cold day when one speaks in relative terms. The effect of this persistence was apparently minimized in this study because the MSN January data (1955–59) showed just ten cases when two successive days were in the same type; MSP January data had nine cases, and MSN July data had six cases with two successive days in the same class and two cases with three successive days in the same class. By minimizing this bias, the validity of the resulting weather types is greatly enhanced.

Clearly, figures 1 and 4 are general models based on limited data samples and their actual operational value may be limited. If more data were applied and these models verified, however, the forecaster could apply them along with the statistical properties of the variables to operational forecasts. In using this approach, however, the forecaster must incorporate his synoptic knowledge and experience to arrive at the best predicted values. The station may experience a great variety of winds and still remain in the same weather type while other types may show wide dispersion in some other parameter. The fact that considerable dispersion occurs within a type does not mean that the days are improperly typed; rather, it means that those variables within a type which display wide dispersion are least important in describing that type. The following simple example illustrates how the forecaster must use his additional information and knowledge to supplement a model of this kind and its accompanying tabular data for operational forecasting. In type A (MSN January), five of the eight midnight temperatures were in the $-3°$ to $0°$ F. range and the remaining three were $8°$ F. The experienced forecaster would surmise that the proper temperature forecast would depend on the snow cover. In checking the condition of the ground on these dates, it was found that the three warm nights had 1 in. or less of snow on the ground with bare spots, while the five cold nights had from 2 to 4 in. of snow on the ground.

The indicated frequencies for each type may be questioned because of the limited sample. Those types containing only two cases may have evolved randomly and only a study using a sufficiently representative data sample will determine whether these types are real or not.

The analyses of variance performed on the mathematical model lend credence to the validity of the model, but they do not represent true checks on the model. It

was impossible to compute the "pure" error in this experiment because this was not a "controlled (or manipulative)" experiment; and because of the order of the matrices involved, it would have been an overwhelming task to graphically compare $\mathbf{y}-\hat{\mathbf{y}}$ vs. $\hat{\mathbf{y}}$. The model did pass the most critical test of all, however, since this experiment provided physically logical results.

In summary, it has been shown that (1) observed weather elements can be expressed in a smaller number of independent elements and these new elements agree with our knowledge of dynamics, (2) these new variables may be grouped into a linear mathematical model for each day and these days may be classified into weather types that are synoptically reasonable, and (3) the distribution of these types relative to the usual array of Highs and Lows falls into patterns that are similar to satellite cloud models and photographs. Clearly there is a need to apply the methods used here to other times of the year and to other geographical areas. The limited sample used here, and the optional ways available to describe a day meteorologically, suggest that one should interpret the foregoing results as those of a "pilot" study.

REFERENCES

1. E. J. Aubert, I. A. Lund, and A. Thomasell, Jr., "Some Objective Six-Hour Predictions Prepared by Statistical Methods," *Journal of Meteorology*, vol. 16, No. 4, Aug. 1959, pp. 436–446.
2. C. E. P. Brooks and N. Carruthers, *Handbook of Statistical Methods in Meteorology*, Meteorological Office, Air Ministry, London, 1953, 413 pp.
3. R. D. Elliott and J. R. Thompson, *Relationships Between TIROS Cloud Patterns and Air Mass (Wind and Thermal) Structure*, Aerometric Research, Inc., Santa Barbara Airport, Goleta, Calif., Sept. 1965.
4. M. Grimmer, "The Space-Filtering of Monthly Surface Temperature Anomaly Data in Terms of Patterns Using Empirical Orthogonal Functions," *Quarterly Journal of the Royal Meteorological Society*, vol. 89, No. 381, July 1963, pp. 395–408.
5. H. H. Harman, *Modern Factor Analysis*, University of Chicago Press, Chicago, Ill., 1960.
6. I. A. Lund, "Map Pattern Classification by Statistical Methods," *Journal of Applied Meteorology*, vol. 2, No. 1, Feb. 1963, pp. 56–65.
7. M. D. Shulman and R. A. Bryson, "A Statistical Study of Dendroclimatic Relationships in South Central Wisconsin," *Journal of Applied Meteorology*, vol. 4, No. 1, Feb. 1965, pp. 107–111.
8. D. Steiner, "A Multivariate Statistical Approach to Climatic Regionalization and Classification," *Tijdschrift van het Koninklijk Nederlandsch Aardrijkskundig Genootschap*, vol. 82, No. 4, 1965.
9. U.S. Weather Bureau, *Daily Series Synoptic Weather Maps, Part I, Northern Hemisphere Sea Level and 500 Millibar Charts*, 1954–1959.
10. R. M. White, D. S. Cooley, R. C. Derby, and F. A. Seaver, "The Development of Efficient Linear Statistical Operators for the Prediction of Sea Level Pressure," *Journal of Meteorology*, vol. 15, No. 5, Oct. 1958, pp. 426–434.

[*Received August 12, 1966; revised October 12, 1966*]